On Gaussian Process Models for High-Dimensional Geostatistical Datasets

Sudipto Banerjee

Joint work with Abhirup Datta, Andrew O. Finley and Alan E. Gelfand

University of California, Los Angeles, USA

May 14, 2015

Space debris

U.S. Forest biomass data

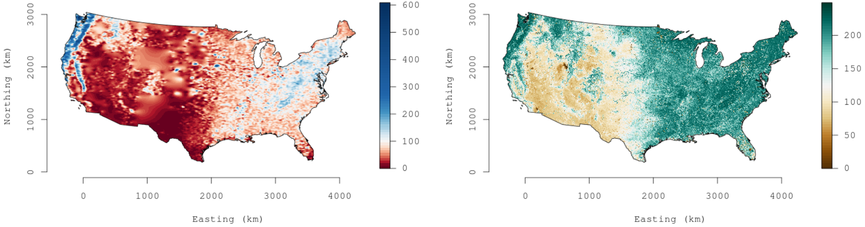


Figure: Observed biomass (left) and NDVI (right)

- Forest biomass data collected between 1999 and 2006 at 114,371 plots
- Normalized Difference Vegetation Index (NDVI) calculated in July 2006
- NDVI is a measure of greenness and is used as a covariate in Forest Biomass Regression Models

Non Spatial Model

Model

$$Biomass = \beta_0 + \beta_1 NDVI + error, \quad \hat{\beta}_0 = 1.043, \; \hat{\beta}_1 = 0.0093$$

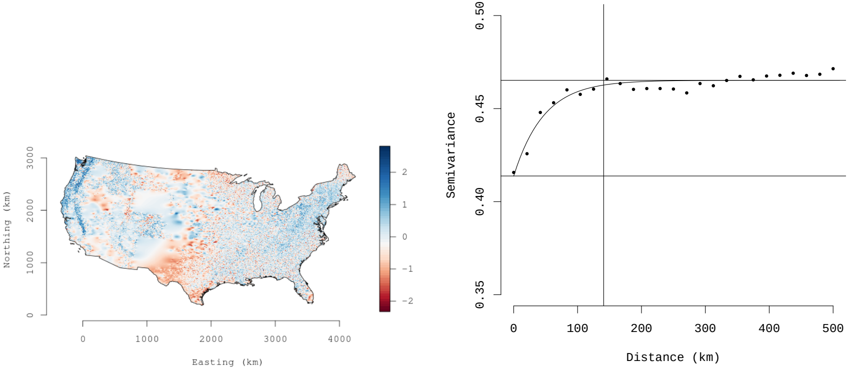


Figure: Heat map (left) and variogram (right) of residuals reflecting spatial correlation

Spatially-varying regression models

- $Y(s) = \beta_0(s) + \beta_1(s)X(s) + e(s)$

- Produce maps for intercept and slope:

$$\left\{\beta_0(s) : s \in D \subset \Re^d\right\} \quad \text{and} \quad \left\{\beta_1(s) : s \in D \subset \Re^d\right\}$$

- This would be rich: understand spatially-varying impact of predictors on outcome.

- Model-based predictions: $Y(s_0) \mid \{y(s_1), y(s_2), \ldots, y(s_n)\}$.

Gaussian (spatial) process

- $\{w(s) : s \in D \subset \Re^d\} \sim GP(0, K_\theta(s,t))$ implies

  $$w = (w(s_1), w(s_2), \ldots, w(s_n))^\top \sim N(0, K_\theta)$$

  for every finite set of points $s_1, s_2, \ldots, s_n$.

- $K_\theta = \{K_\theta(s_i, s_j)\}$ is a spatial variance-covariance matrix

- Stationary: $K_\theta(s,t) = K_\theta(t-s)$. Isotropy: $K_\theta(s,t) = K_\theta(\|t-s\|)$.

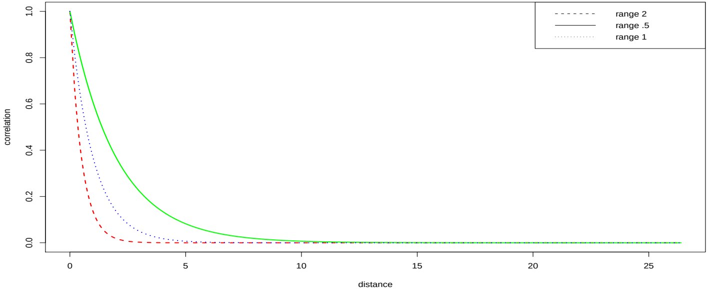
- Bochner: Covariance function $\Leftrightarrow$ characteristic function.

Matérn covariance:

$$K_\theta(s,t) = \frac{\sigma^2}{2^{\phi_2-1}\Gamma(\phi_2)}(\|t-s\|\phi_1)^{\phi_2}\kappa_{\phi_2}(\|t-s\|;\phi_1)$$

ϕ_1 → controls how fast correlation decays

ϕ_2 → controls smoothness of the spatial surface



Hierarchical Gaussian process models

Full rank model

- $T = \{t_1, t_2, \ldots, t_n\}$ are locations where data is observed
- $y(t_i)$ is outcome at the i^{th} location, $y = (y(t_1), y(t_2), \ldots, y(t_n))^\top$
- $y = X\beta + Zw + \epsilon$, $\epsilon \sim N(0, \tau^2 I)$
- $w = (w(t_1), w(t_2), \ldots, w(t_n))^\top$ are spatial random effects
- $w \sim N(0, K_\theta)$, K_θ is a valid spatial covariance matrix
- Priors on $\{\beta, \tau^2, \theta\}$

Computation issues

- Storage: n^2 pairwise distances to compute K_θ
- K_θ is dense; solve $K_\theta x = b$ *and* need $\det(K_\theta)$
- Complexity: roughly $O(n^3)$ flops; computationally infeasible for large datasets

Burgeoning literature on spatial big data

- Low-rank approaches (Wahba, 1990; Higdon, 2002; Kamman & Wand, 2003; Paciorek, 2007; Rasmussen & Williams, 2006; Stein 2007, 2008; Cressie & Johannesson, 2008; Banerjee et al., 2008; 2010; Gramacy & Lee 2008; Sang et al., 2011; Lemos et al., 2011; Guhaniyogi et al., 2011, 2013; Salazar et al., 2013)
- Covariance tapering (Furrer et al. 2006; Zhang and Du, 2007; Du et al. 2009; Kaufman et al., 2009)
- Spectral domain: (Fuentes 2007; Paciorek, 2007)
- Approximation using GMRFs: `INLA` (Rue et al. 2009; Lindgren et al., 2011)
- Nearest-neighbor models (processes) (Vecchia 1988; Stein et al. 2004; Gramacy et al. 2014; Stroud et al 2014; Datta et al., 2015)

Low-rank models: hierarchical approach

$$N(w^* \,|\, 0, K_\theta^*) \times N(y \,|\, B_\theta w^*, D)$$

- y is $n \times 1$ and n is large
- w^* is $r \times 1$, where $r << n$; so K_θ^* is $r \times r$
- B_θ is $n \times r$ is a matrix of "basis" functions
- D is $n \times n$, but easy to invert (e.g. diagonal)
- Derive $\text{var}(y)$ (or $\text{var}(w^* \,|\, y)$) in two ways to obtain

$$(D + B_\theta K_\theta^* B_\theta^\top)^{-1} = D^{-1} - D^{-1} B_\theta (K_\theta^{*-1} + B_\theta^\top D^{-1} B_\theta)^{-1} B_\theta^\top D^{-1} \ .$$

- This is the famous Sherman-Woodbury-Morrison formula.
- Modeling: specifying w^* and B_θ.

Gaussian predictive process (Banerjee et al., *JRSS-B*, 2008)

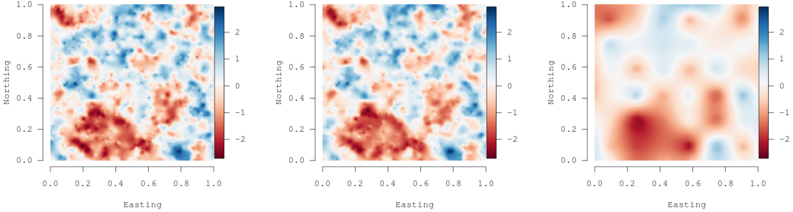
- Start with a parent Gaussian process $w(s) \sim GP(0, K_\theta(\cdot, \cdot))$
- Fix a set of "knots" $s_1, s_2, \ldots, s_r$, and let $K_\theta^* = \{K_\theta(s_i, s_j)\}$
- Then, $w^* = (w(s_1), w(s_2), \ldots, w(s_r))^\top \sim N(0, K_\theta^*)$

Gaussian predictive process (Banerjee et al., *JRSS-B*, 2008)

- Start with a parent Gaussian process $w(s) \sim GP(0, K_\theta(\cdot,\cdot))$
- Fix a set of "knots" $s_1, s_2, \ldots, s_r$, and let $K_\theta^* = \{K_\theta(s_i, s_j)\}$
- Then, $w^* = (w(s_1), w(s_2), \ldots, w(s_r))^\top \sim N(0, K_\theta^*)$
- Predictive process: $\tilde{w}(s) = \mathrm{E}[w(s) \,|\, w^*] = b_\theta(s)^\top w^*$
- Orthogonal decomposition:

$$\mathrm{var}\{w(s)\} = \mathrm{var}\{\tilde{w}(s)\} + \mathrm{var}\{w(s) - \tilde{w}(s)\}$$

- Approximate *residual* process with a *sparse* process (Sang et al. 2011)



(a) True w

(b) Full GP

(c) PPGP 64 knots

Figure: Comparing full GP vs low-rank GP with 2000 locations

Sparse Gaussian Processes

- Introduce (auxiliary) random effects to achieve computational benefits.
- Let $\mathcal{S} = \{s_1, s_2, \ldots, s_k\}$ be a “reference” set of points.

$$\text{Spatial random effects: } (w(s_1), w(s_2), \ldots, w(s_k))^\top \sim N(0, \tilde{K}_\theta) \; ,$$

$$\text{Spatial process: } w(t) = \sum_{i=1}^{k} a_i(t) w(s_i) + \eta(t) \; .$$

1. Example: $\eta(t) \overset{ind}{\sim} N(0, \tau^2(t))$.
2. Example: $a_i(t) \neq 0$ ONLY IF t is a “neighbor” of s_i.

- Three pieces to the puzzle:

1. How do we construct $\tilde{K}_\theta^{-1}$ to be sparse and $\det(\tilde{K}_\theta)$ to be cheap?
2. How do we define “neighbors” for arbitrary points t?
3. How do we choose nonzero $a_i(t)$’s? Ensure good approx. to full GP?

Simple method of introducing sparsity (e.g. graphical models)

- Write a joint density $p(w) = p(w_1, w_2, \ldots, w_n)$ as:

$$p(w_1)p(w_2 \mid w_1)p(w_3 \mid w_1, w_2) \cdots p(w_n \mid w_1, w_2, \ldots, w_{n-1})$$

- Example: For Gaussian distributions:

$$\begin{aligned} w_1 &= 0 + \eta_1; \\ w_i &= a_{i1}w_1 + a_{i2}w_2 + \cdots + a_{i,i-1}w_{i-1} + \eta_i; \quad i = 2, 3, \ldots, n \\ \Longrightarrow w &= Aw + \eta; \quad \eta \sim N(0, D) \end{aligned}$$

- Making some $a_{ij} = 0$ introduces conditional independence

- Equivalent to $w \sim N(0, K_\theta)$ and $\texttt{chol}(K_\theta^{-1}) = LDL^\top$, then additional zeroes in lower-triangular L.

Sparse likelihood approximations (Vecchia, 1988; Stein et al., 2004)

- With $w_i \equiv w(s_i)$, write a GP joint density $p(w) = p(w_1, w_2, \ldots, w_n)$ as:

$$p(w_1)p(w_2 \,|\, w_1)p(w_3 \,|\, w_1, w_2) \cdots p(w_n \,|\, w_1, w_2, \ldots, w_{n-1})$$

- Use “screening effect” to impose conditional independence and obtain:

$$\tilde{p}(w) = p(w_1)p(w_2 \,|\, w_1)p(w_3 \,|\, w_1, w_2)p(w_4 \,|\, w_1, w_3) \cdots p(w_n \,|\, w_{i_n}, w_{j_n})$$

- If $w \sim N(0, K_\theta)$, then $\tilde{p}(w) = N(w \,|\, 0, \tilde{K}_\theta)$
- $\tilde{K}_\theta^{-1}$ is *sparser* than K_θ^{-1}.

Sparse precision matrices

Two crucial facts

1. $\tilde{p}(w)$ is a valid joint density from the model $w \sim N(0, \tilde{K}_\theta)$
2. $\tilde{K}_\theta^{-1}$ depends on K_θ and is sparse with at most nm^2 non-zero entries

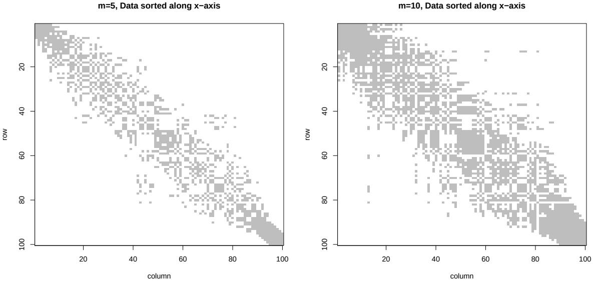


Figure: Sparse precision matrices from neighbor-based approximation

Extension to a Nearest-neighbor GP (Datta et al., *JASA*, 2015)

- Fix any “reference” set $S = \{s_1, s_2, \ldots, s_k\}$

$$N(s_i) = \begin{cases} \text{empty set for } i = 1 \\ \{s_1, s_2, \ldots, s_{i-1}\} \text{ for } 2 \le i \le m \\ m \text{ nearest neighbors of } s_i \text{ among } \{s_1, s_2, \ldots, s_{i-1}\} \text{ for } i > m \end{cases}$$

- Model $w_S \sim N(0, \tilde{K}_\theta)$ (“Vecchia prior”)
- For any t outside S, define $N(t)$ as the set of m-nearest neighbors of t in S
- Construct $w(t) = \sum_{i=1}^{k} a_i(t) w(s_i) + \eta(t)$ with $a_i(t) = 0$ if $s_i \notin N(t)$.
- Nonzero $a_i(t)$’s are specified according to $p(w(t) \mid w_{N(t)})$.

- For $T = \{t_1, t_2, \ldots, t_n\}$ outside S, we define

$$\tilde{p}(w_T \,|\, w_S) = \prod_{i=1}^{n} p(w(t_i) \,|\, w_{N(t_i)}) \ .$$

- Generalize to any finite T as follows:

$$\tilde{p}(w_T) = \int \tilde{p}(w_S)\tilde{p}(w_{T\setminus S} \,|\, w_S) \prod_{\{i \,|\, s_i \in S\setminus T\}} d(w(s_i))$$

- Example: Model $\tilde{p}(w_S)\tilde{p}(w_T \,|\, w_S) = N(w_S \,|\, 0, \tilde{K}_\theta) \times N(w_T \,|\, A_T w_S, D_T)$

- A very convenient choice in practice: $S = T$, i.e., take set of observed locations as reference set.

Hierarchical NNGP model

NNGP used as a sparsity inducing prior for hierarchical models.

Likelihood

$$\begin{aligned} N(y \,|\, X\beta + Zw_T, \tau^2 I) \times N(w_T \,|\, A_T w_S, D_T) \times N(w_S \,|\, 0, \tilde{K}_\theta) \\ \times \quad N(\beta \,|\, \mu_\beta, V_\beta) \times IG(\tau^2 \,|\, a_\tau, b_\tau) \times \pi(\theta) \end{aligned}$$

Gibbs' sampler

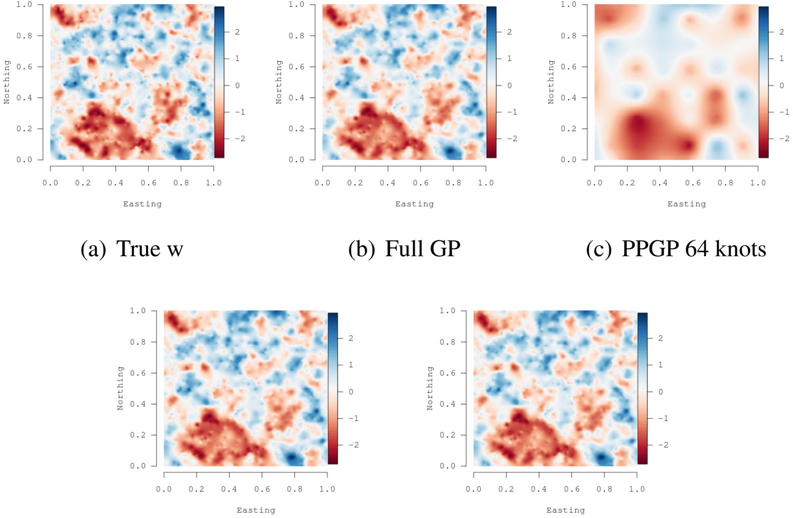
- Conjugate full conditionals for β, τ^2
- Sequential updates for full conditional of $w(t_i)$'s
- Metropolis step for updating θ

Storage and computation

- Never needs to store $n \times n$ distance matrix. Stores n small $m \times m$ matrices
- Total flop count per iteration of Gibbs' sampler is $O(nm^3)$ i.e linear in n
- Scalable to massive datasets

Simulation experiments

- 2500 locations on a unit square
- $y(t_i) = \beta_0 + \beta_1 X(t_i) + w(t_i) + \epsilon(t_i)$
- Single covariate generated from $N(0, 1)$
- Spatial effects generated from $GP(0, \sigma^2 R(\nu, \phi))$
- $R(\nu, \phi)$ is Matern correlation function with smoothness ν and decay ϕ
- Candidate models: Full GP, Low rank GP (PPGP) with 64 knots and NNGP



(a) True w

(b) Full GP

(c) PPGP 64 knots

(d) NNGP, $m = 10$

(e) NNGP, $m = 20$

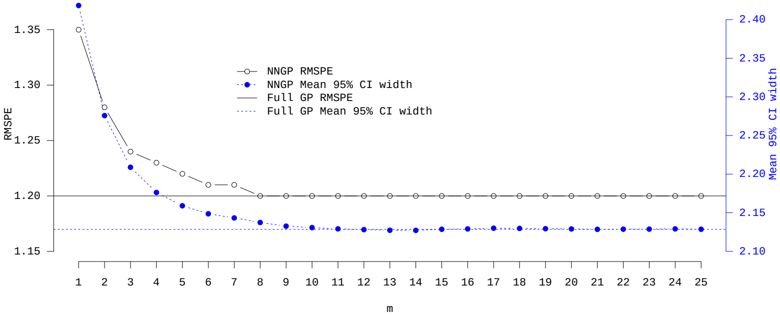


Figure: Choice of *m* in NNGP models: Out-of-sample Root Mean Squared Prediction Error (RMSPE) and mean width between the upper and lower 95% posterior predictive credible intervals for a range of *m* for the univariate synthetic data analysis

Table: Univariate synthetic data analysis

	True	NNGP $m = 10$	 $m = 20$	Predictive Process 64 knots	Full Gaussian Process
β_0	1	1.00 (0.62, 1.31)	1.03 (0.65, 1.34)	1.30 (0.54, 2.03)	1.03 (0.69, 1.34)
β_1	5	5.01 (4.99, 5.03)	5.01 (4.99, 5.03)	5.03 (4.99, 5.06)	5.01 (4.99, 5.03)
σ^2	1	0.96 (0.78, 1.23)	0.94 (0.77, 1.20)	1.29 (0.96, 2.00)	0.94 (0.76, 1.23)
τ^2	0.1	0.10 (0.08, 0.13)	0.10 (0.08, 0.13)	0.08 (0.04, 0.13)	0.10 (0.08, 0.12)
ϕ	12	12.93 (9.70, 16.77)	13.36 (9.99, 17.15)	5.61 (3.48, 8.09)	13.52 (9.92, 17.50)
G (Goodness of fit)	–	77.84	76.40	1075.63	74.80
P (Penalization)	–	340.40	337.88	200.39	333.27
D (G+P)	–	418.24	414.28	1276.03	408.08
RMSPE	–	1.2	1.2	1.68	1.2
Run time (Minutes)	–	14.40	46.47	43.36	560.31

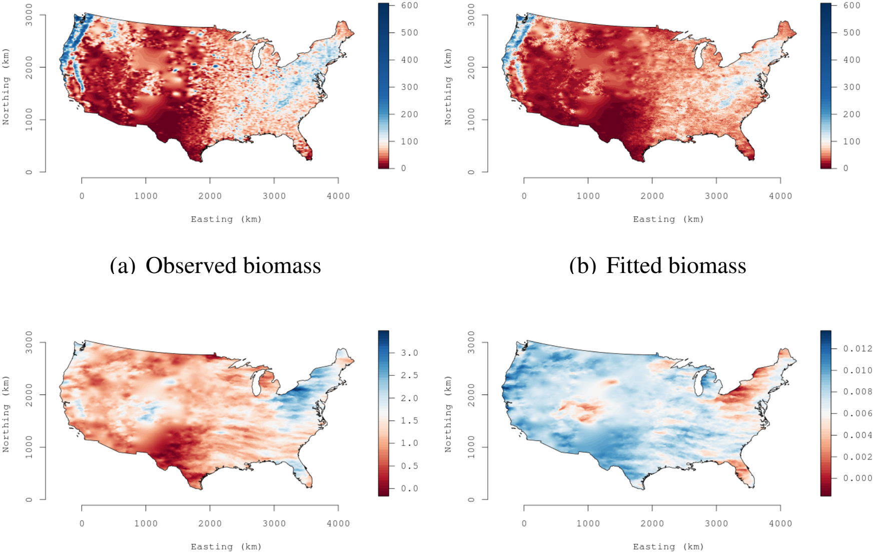
- Parameter estimates for all models are similar
- NNGP performs at par with Full GP, PPGP performs worse
- NNGP yields huge computational gains

Back to the Forest biomass dataset

- Number of spatial locations: $n = 114,371$
- Full GP and PPGP storage requirements $\gg$ 38 gigabytes available
- We use a hierarchical spatially varying coefficients NNGP model

Model

- $Biomass(t) = (\beta_0 + \beta_0(t)) + (\beta_1 + \beta_1(t))NDVI(t) + \epsilon(t)$
- $w(t) = (\beta_0(t), \beta_1(t))^\top \sim$ Bivariate $NNGP(0, \tilde{K}_\theta(\cdot))$, $m = 5$
- Full inferential output: 46 hrs



(a) Observed biomass

(b) Fitted biomass

(c) $\beta_0(t)$

(d) $\beta_{NDVI}(t)$

Conclusions

- Unified platform for estimation, prediction and model comparison
- Easily extends to multivariate and spatial-temporal processes
- Posterior predictions, recovery of latent spatial surfaces
- Superior performance, massive computation and storage gains over existing models
- Possible extension to spatial GLMs

Thank you!